Nano GPT

H   i   !

Token Embedding Table

0,1,2

Positional Embedding Table

+

T (3) x n_embd (384)

Single head of attention

linear n_embd -> head_size (6)

linear n_embd -> head_size (6)

linear n_embed -> head_size (6)

Query H

Query i

Query !

Key H

Key i

Key !

Value H

Value i

Value !

q @ transpose(k)

masked fill tril

(T x head_size)

softmax

w HH		
w Hi	w ii	
w H!	w i!	w !!

head size 6

(T x T)

@

(T x head_size)

head out H

head out i

head out !

concat along token dimension

Multi Headed Attention (T*n_head x head_size)

Linear (n_head x head_size -> n_embd)

output
Txn_embd

Mutliheaded Attention
Input: T x n_embed
Output:Tx n_embed

Feedforward (n_embd -> n_embd)

Linear (n_embd -> 4 * n_embd)

Relu

Linear (4 * n_embd > n_embd)

Block

T x n_embed → Layer Norm —Txn_embed→ Multiheaded Attention —Txn_embed→ Layer Norm → T x n_embed

GPT Model

T x n_embed → Block —Txn_embed→ Block —Txn_embed→ Layer Norm → Linear n_embed -> vocab_size → vocab_size

n_layers number of sequential blocks

logits